# Data Pre-Processing

# importing data

```
Warning: Looks like you're using an outdated `kagglehub` version, please consider
updating (latest version: 0.3.10)
Path to dataset files: /home/leoadmin/.cache/kagglehub/datasets/technika148/footba
ll-database/versions/1
```

```
Files and directories in ' /home/leoadmin/.cache/kagglehub/datasets/technika148/fo
otball-database/versions/1 ' :
['leagues.csv', 'teamstats.csv', 'teams.csv', 'shots.csv', 'players.csv', 'appeara
nces.csv', 'games.csv']
```

# Creating data frames from data

```
Loaded DataFrames: dict_keys(['leagues', 'teamstats', 'teams', 'shots', 'players',
'appearances', 'games'])
```

| | gameID | playerID | goals | ownGoals | shots | xGoals | xGoalsChain | xGoalsBuildup | assists | keyP |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 560 | 0 | 0 | 0 | 0.0 | 0.000000 | 0.000000 | 0 | |
| **1** | 81 | 557 | 0 | 0 | 0 | 0.0 | 0.106513 | 0.106513 | 0 | |
| **2** | 81 | 548 | 0 | 0 | 0 | 0.0 | 0.127738 | 0.127738 | 0 | |
| **3** | 81 | 628 | 0 | 0 | 0 | 0.0 | 0.106513 | 0.106513 | 0 | |
| **4** | 81 | 1006 | 0 | 0 | 0 | 0.0 | 0.021225 | 0.021225 | 0 | |

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 356513 entries, 0 to 356512
Data columns (total 19 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   gameID         356513 non-null  int64
 1   playerID       356513 non-null  int64
 2   goals          356513 non-null  int64
 3   ownGoals       356513 non-null  int64
 4   shots          356513 non-null  int64
 5   xGoals         356513 non-null  float64
 6   xGoalsChain    356513 non-null  float64
 7   xGoalsBuildup  356513 non-null  float64
 8   assists        356513 non-null  int64
 9   keyPasses      356513 non-null  int64
 10  xAssists       356513 non-null  float64
 11  position       356513 non-null  object
 12  positionOrder  356513 non-null  int64
 13  yellowCard     356513 non-null  int64
 14  redCard        356513 non-null  int64
 15  time           356513 non-null  int64
 16  substituteIn   356513 non-null  int64
 17  substituteOut  356513 non-null  int64
 18  leagueID       356513 non-null  int64
dtypes: float64(4), int64(14), object(1)
memory usage: 51.7+ MB
None
```

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals | home |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 | |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 | |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 | |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 | |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 | |

5 rows × 34 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12680 entries, 0 to 12679
Data columns (total 34 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   gameID             12680 non-null  int64  
 1   leagueID           12680 non-null  int64  
 2   season             12680 non-null  int64  
 3   date               12680 non-null  object 
 4   homeTeamID         12680 non-null  int64  
 5   awayTeamID         12680 non-null  int64  
 6   homeGoals          12680 non-null  int64  
 7   awayGoals          12680 non-null  int64  
 8   homeProbability    12680 non-null  float64
 9   drawProbability    12680 non-null  float64
 10  awayProbability    12680 non-null  float64
 11  homeGoalsHalfTime  12680 non-null  int64  
 12  awayGoalsHalfTime  12680 non-null  int64  
 13  B365H              12675 non-null  float64
 14  B365D              12675 non-null  float64
 15  B365A              12675 non-null  float64
 16  BWH                12677 non-null  float64
 17  BWD                12677 non-null  float64
 18  BWA                12677 non-null  float64
 19  IWH                12662 non-null  float64
 20  IWD                12662 non-null  float64
 21  IWA                12662 non-null  float64
 22  PSH                12660 non-null  float64
 23  PSD                12660 non-null  float64
 24  PSA                12660 non-null  float64
 25  WHH                12674 non-null  float64
 26  WHD                12674 non-null  float64
 27  WHA                12674 non-null  float64
 28  VCH                12676 non-null  float64
 29  VCD                12676 non-null  float64
 30  VCA                12676 non-null  float64
 31  PSCH               12678 non-null  float64
 32  PSCD               12678 non-null  float64
 33  PSCA               12678 non-null  float64
dtypes: float64(24), int64(9), object(1)
memory usage: 3.3+ MB
None
```

| | leagueID | name | understatNotation |
|---|---|---|---|
| **0** | 1 | Premier League | EPL |
| **1** | 2 | Serie A | Serie_A |
| **2** | 3 | Bundesliga | Bundesliga |
| **3** | 4 | La Liga | La_liga |
| **4** | 5 | Ligue 1 | Ligue_1 |

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 3 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   leagueID           5 non-null      int64
 1   name               5 non-null      object
 2   understatNotation  5 non-null      object
dtypes: int64(1), object(2)
memory usage: 248.0+ bytes
None
```

| | playerID | name |
|---|---|---|
| **0** | 560 | Sergio Romero |
| **1** | 557 | Matteo Darmian |
| **2** | 548 | Daley Blind |
| **3** | 628 | Chris Smalling |
| **4** | 1006 | Luke Shaw |

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7659 entries, 0 to 7658
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   playerID  7659 non-null   int64
 1   name      7659 non-null   object
dtypes: int64(1), object(1)
memory usage: 119.8+ KB
None
```

| | teamID | name |
|---|---|---|
| **0** | 71 | Aston Villa |
| **1** | 72 | Everton |
| **2** | 74 | Southampton |
| **3** | 75 | Leicester |
| **4** | 76 | West Bromwich Albion |

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 146 entries, 0 to 145
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   teamID  146 non-null    int64
 1   name    146 non-null    object
dtypes: int64(1), object(1)
memory usage: 2.4+ KB
None
```

| | gameID | shooterID | assisterID | minute | situation | lastAction | shotType | shotResult | |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 554 | NaN | 27 | DirectFreekick | Standard | LeftFoot | BlockedShot | 0. |
| **1** | 81 | 555 | 631.0 | 27 | SetPiece | Pass | RightFoot | BlockedShot | 0. |
| **2** | 81 | 554 | 629.0 | 35 | OpenPlay | Pass | LeftFoot | BlockedShot | 0. |
| **3** | 81 | 554 | NaN | 35 | OpenPlay | Tackle | LeftFoot | MissedShots | 0. |
| **4** | 81 | 555 | 654.0 | 40 | OpenPlay | BallRecovery | RightFoot | BlockedShot | 0. |

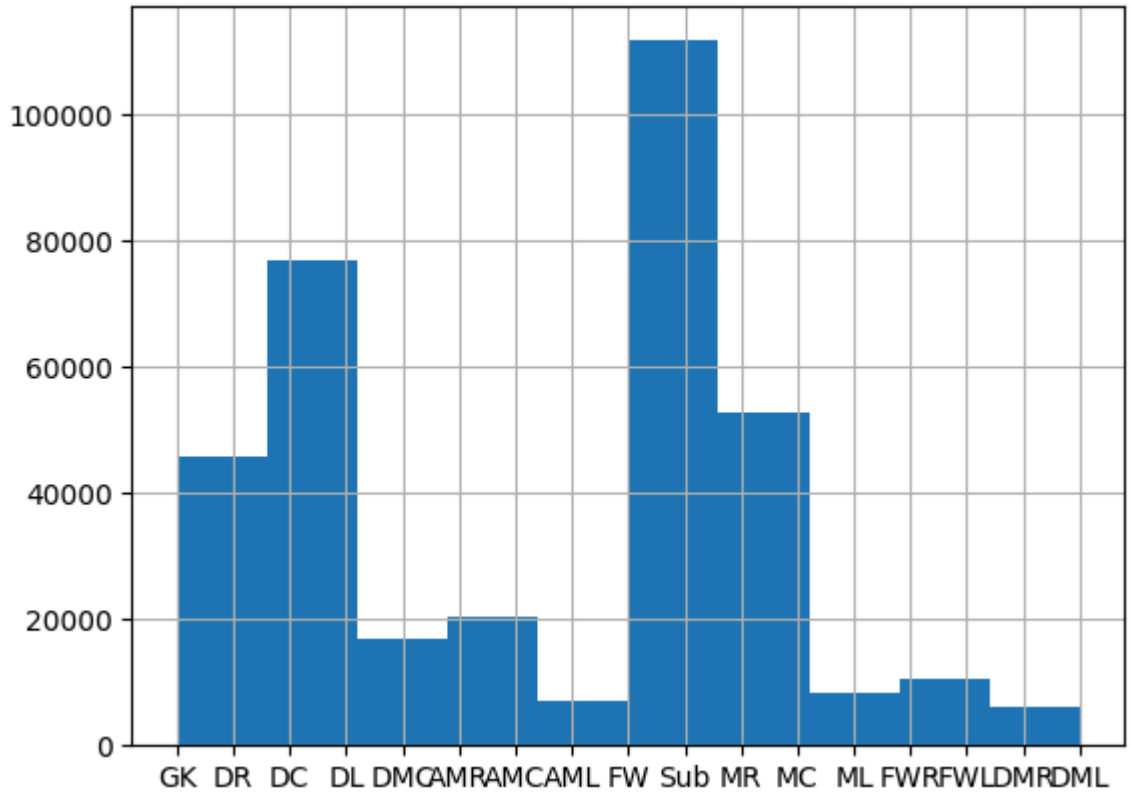
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 324543 entries, 0 to 324542
Data columns (total 11 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   gameID      324543 non-null  int64
 1   shooterID   324543 non-null  int64
 2   assisterID  240199 non-null  float64
 3   minute      324543 non-null  int64
 4   situation   324543 non-null  object
 5   lastAction  324543 non-null  object
 6   shotType    324543 non-null  object
 7   shotResult  324543 non-null  object
 8   xGoal       324543 non-null  float64
 9   positionX   324543 non-null  float64
 10  positionY   324543 non-null  float64
dtypes: float64(4), int64(3), object(4)
memory usage: 27.2+ MB
None
```

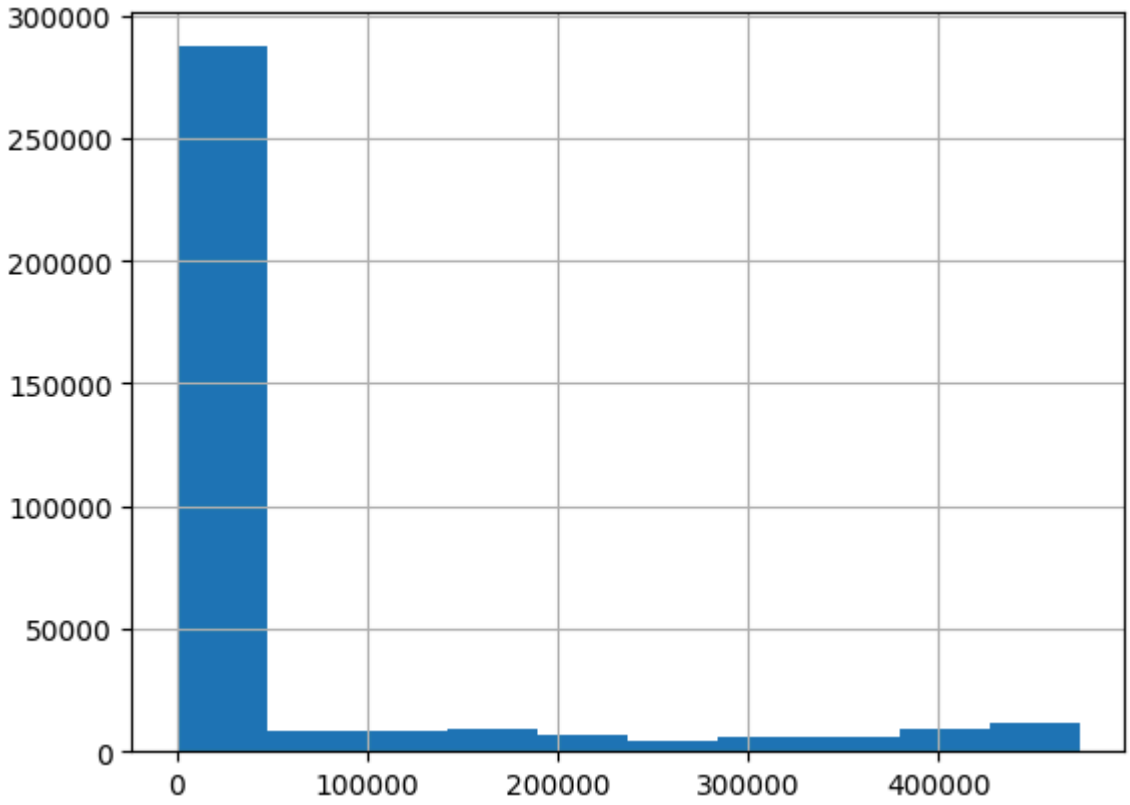
| | gameID | teamID | season | date | location | goals | xGoals | shots | shotsOnTarget | deep | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 89 | 2015 | 2015-08-08 15:45:00 | h | 1 | 0.627539 | 9 | 1 | 4 | 13 |
| **1** | 81 | 82 | 2015 | 2015-08-08 15:45:00 | a | 0 | 0.674600 | 9 | 4 | 10 | 8 |
| **2** | 82 | 73 | 2015 | 2015-08-08 18:00:00 | h | 0 | 0.876106 | 11 | 2 | 11 | 6 |
| **3** | 82 | 71 | 2015 | 2015-08-08 18:00:00 | a | 1 | 0.782253 | 7 | 3 | 2 | 11 |
| **4** | 83 | 72 | 2015 | 2015-08-08 18:00:00 | h | 2 | 0.604226 | 10 | 5 | 5 | 6 |

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25360 entries, 0 to 25359
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   gameID         25360 non-null  int64
 1   teamID         25360 non-null  int64
 2   season         25360 non-null  int64
 3   date           25360 non-null  object
 4   location       25360 non-null  object
 5   goals          25360 non-null  int64
 6   xGoals         25360 non-null  float64
 7   shots          25360 non-null  int64
 8   shotsOnTarget  25360 non-null  int64
 9   deep           25360 non-null  int64
 10  ppda           25360 non-null  float64
 11  fouls          25360 non-null  int64
 12  corners        25360 non-null  int64
 13  yellowCards    25359 non-null  float64
 14  redCards       25360 non-null  int64
 15  result         25360 non-null  object
dtypes: float64(3), int64(10), object(3)
memory usage: 3.1+ MB
None

17    77563
3     56289
9     44453
15    34094
1     25358
4     20559
2     20556
7     16903
12    13542
10     8390
8      8389
11     6923
13     6923
14     5211
16     5211
5      3075
6      3074
Name: positionOrder, dtype: int64
```

# Having a look into the dfs


100000
80000
60000
40000
20000
0
GK DR DC DL DMCAMRAMCAML FW Sub MR MC ML FWRFWLDMRDML


300000
250000
200000
150000
100000
50000
0
0
100000
200000
300000
400000

```
0         278954
316647         1
317386         1
317387         1
317382         1
           ...
33206          1
33209          1
33186          1
33194          1
474074         1
Name: substituteOut, Length: 77560, dtype: int64
```

```
0         278954
316652         1
317393         1
317391         1
317392         1
           ...
33210          1
33212          1
33197          1
33196          1
474075         1
Name: substituteIn, Length: 77560, dtype: int64
```

| | gameID | playerID | substituteIn | substituteOut | time |
|---|---|---|---|---|---|
| **0** | 81 | 560 | 0 | 0 | 90 |
| **1** | 81 | 557 | 222605 | 0 | 82 |
| **2** | 81 | 548 | 0 | 0 | 90 |
| **3** | 81 | 628 | 0 | 0 | 90 |
| **4** | 81 | 1006 | 0 | 0 | 90 |
| **...** | ... | ... | ... | ... | ... |
| **356508** | 16135 | 3509 | 0 | 0 | 90 |
| **356509** | 16135 | 4882 | 0 | 0 | 90 |
| **356510** | 16135 | 5786 | 0 | 0 | 90 |
| **356511** | 16135 | 8997 | 474075 | 0 | 78 |
| **356512** | 16135 | 5762 | 0 | 474074 | 12 |

356513 rows × 5 columns

changing the logic of substitute in/out and repalcing it to be binary

| | gameID | playerID | goals | ownGoals | shots | xGoals | xGoalsChain | xGoalsBuildup | assists | ke |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 560 | 0 | 0 | 0 | 0.000000 | 0.000000 | 0.000000 | 0 | |
| **1** | 81 | 557 | 0 | 0 | 0 | 0.000000 | 0.106513 | 0.106513 | 0 | |
| **2** | 81 | 548 | 0 | 0 | 0 | 0.000000 | 0.127738 | 0.127738 | 0 | |
| **3** | 81 | 628 | 0 | 0 | 0 | 0.000000 | 0.106513 | 0.106513 | 0 | |
| **4** | 81 | 1006 | 0 | 0 | 0 | 0.000000 | 0.021225 | 0.021225 | 0 | |
| **5** | 81 | 551 | 0 | 0 | 0 | 0.000000 | 0.163670 | 0.163670 | 0 | |
| **6** | 81 | 654 | 0 | 0 | 0 | 0.000000 | 0.035742 | 0.000000 | 0 | |
| **7** | 81 | 554 | 0 | 0 | 3 | 0.253645 | 0.255811 | 0.106513 | 0 | |
| **8** | 81 | 555 | 0 | 0 | 3 | 0.121309 | 0.056967 | 0.000000 | 0 | |
| **9** | 81 | 631 | 0 | 0 | 1 | 0.103004 | 0.124229 | 0.021225 | 0 | |
| **10** | 81 | 629 | 0 | 0 | 2 | 0.149581 | 0.184895 | 0.127738 | 0 | |
| **11** | 81 | 552 | 0 | 0 | 0 | 0.000000 | 0.106513 | 0.106513 | 0 | |
| **12** | 81 | 627 | 0 | 0 | 0 | 0.000000 | 0.000000 | 0.000000 | 0 | |
| **13** | 81 | 907 | 0 | 0 | 0 | 0.000000 | 0.106513 | 0.106513 | 0 | |
| **14** | 81 | 651 | 0 | 0 | 0 | 0.000000 | 0.000000 | 0.000000 | 0 | |
| **15** | 81 | 638 | 0 | 1 | 1 | 0.073058 | 0.073058 | 0.000000 | 0 | |
| **16** | 81 | 639 | 0 | 0 | 2 | 0.056997 | 0.036491 | 0.000000 | 0 | |
| **17** | 81 | 640 | 0 | 0 | 0 | 0.000000 | 0.109549 | 0.109549 | 0 | |
| **18** | 81 | 660 | 0 | 0 | 0 | 0.000000 | 0.066174 | 0.066174 | 0 | |
| **19** | 81 | 914 | 0 | 0 | 0 | 0.000000 | 0.000000 | 0.000000 | 0 | |
| **20** | 81 | 643 | 0 | 0 | 0 | 0.000000 | 0.073058 | 0.073058 | 0 | |
| **21** | 81 | 642 | 0 | 0 | 0 | 0.000000 | 0.103433 | 0.103433 | 0 | |
| **22** | 81 | 646 | 0 | 0 | 3 | 0.363478 | 0.466911 | 0.000000 | 0 | |
| **23** | 81 | 648 | 0 | 0 | 0 | 0.000000 | 0.297304 | 0.297304 | 0 | |
| **24** | 81 | 647 | 0 | 0 | 3 | 0.182191 | 0.357362 | 0.029683 | 0 | |
| **25** | 81 | 649 | 0 | 0 | 0 | 0.000000 | 0.066174 | 0.066174 | 0 | |
| **26** | 81 | 644 | 0 | 0 | 0 | 0.000000 | 0.029683 | 0.000000 | 0 | |
| **27** | 81 | 645 | 0 | 0 | 0 | 0.000000 | 0.066174 | 0.066174 | 0 | |
| **28** | 82 | 455 | 0 | 0 | 0 | 0.000000 | 0.074331 | 0.074331 | 0 | |
| **29** | 82 | 456 | 0 | 0 | 0 | 0.000000 | 0.239071 | 0.191200 | 0 | |

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 356513 entries, 0 to 356512
Data columns (total 18 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   gameID         356513 non-null  int64
 1   playerID       356513 non-null  int64
 2   goals          356513 non-null  int64
 3   ownGoals       356513 non-null  int64
 4   shots          356513 non-null  int64
 5   xGoals         356513 non-null  float64
 6   xGoalsChain    356513 non-null  float64
 7   xGoalsBuildup  356513 non-null  float64
 8   assists        356513 non-null  int64
 9   keyPasses      356513 non-null  int64
 10  xAssists       356513 non-null  float64
 11  positionOrder  356513 non-null  int64
 12  yellowCard     356513 non-null  int64
 13  redCard        356513 non-null  int64
 14  time           356513 non-null  int64
 15  subOut         356513 non-null  int64
 16  subIn          356513 non-null  int64
 17  leagueID       356513 non-null  int64
dtypes: float64(4), int64(14)
memory usage: 49.0 MB
None
```

extracting odds and probabilities to a separate df, it is not a good feature for modeling

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals | home |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 | |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 | |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 | |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 | |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 | |

| | gameID | homeProbability | drawProbability | awayProbability | B365H | B365D | B365A | BWH | BW |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 0.2843 | 0.3999 | 0.3158 | 1.65 | 4.0 | 6.00 | 1.65 | 4 |
| **1** | 82 | 0.3574 | 0.3500 | 0.2926 | 2.00 | 3.6 | 4.00 | 2.00 | 3 |
| **2** | 83 | 0.2988 | 0.4337 | 0.2675 | 1.70 | 3.9 | 5.50 | 1.70 | 3 |
| **3** | 84 | 0.6422 | 0.2057 | 0.1521 | 1.95 | 3.5 | 4.33 | 2.00 | 3 |
| **4** | 85 | 0.1461 | 0.2159 | 0.6380 | 2.55 | 3.3 | 3.00 | 2.60 | 3 |

5 rows × 25 columns

having a look at the shots df

| | gameID | shooterID | assisterID | minute | situation | lastAction | shotType | shotResu |
|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 554 | NaN | 27 | DirectFreekick | Standard | LeftFoot | BlockedSh |
| **1** | 81 | 555 | 631.0 | 27 | SetPiece | Pass | RightFoot | BlockedSh |
| **2** | 81 | 554 | 629.0 | 35 | OpenPlay | Pass | LeftFoot | BlockedSh |
| **3** | 81 | 554 | NaN | 35 | OpenPlay | Tackle | LeftFoot | MissedSho |
| **4** | 81 | 555 | 654.0 | 40 | OpenPlay | BallRecovery | RightFoot | BlockedSh |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **324538** | 16135 | 6615 | 8651.0 | 19 | SetPiece | Aerial | Head | MissedSho |
| **324539** | 16135 | 6615 | 8651.0 | 54 | SetPiece | Cross | LeftFoot | Go |
| **324540** | 16135 | 3464 | NaN | 70 | OpenPlay | None | LeftFoot | MissedSho |
| **324541** | 16135 | 8651 | 4882.0 | 72 | OpenPlay | Cross | Head | BlockedSh |
| **324542** | 16135 | 8651 | 4882.0 | 85 | OpenPlay | Pass | RightFoot | MissedSho |

324543 rows × 11 columns

there isn't a connection between the team and the shooter so it is complicated to map a shot mapping of each game for each team

```
OpenPlay          237543
FromCorner         47208
SetPiece           21354
DirectFreekick     14451
Penalty             3987
Name: situation, dtype: int64
```

```
Pass               115861
Cross               46175
None                36896
Aerial              23882
Standard            18438
TakeOn              17331
Chipped             16959
Rebound             13735
HeadPass             7997
BallRecovery         7256
Throughball          6459
BallTouch            4995
LayOff               3076
Dispossessed         1780
Tackle                761
Foul                  533
CornerAwarded         464
Interception          424
BlockedPass           370
End                   246
Goal                  204
Challenge             140
Clearance             121
OffsidePass            93
Card                   86
GoodSkill              67
Save                   60
SubstitutionOn         48
FormationChange        16
Start                  14
KeeperPickup           13
Error                  11
Punch                   7
OffsideProvoked         7
ShieldBallOpp           5
KeeperSweeper           4
ChanceMissed            3
PenaltyFaced            2
CrossNotClaimed         2
Smother                 1
SubstitutionOff         1
Name: lastAction, dtype: int64
RightFoot        166121
LeftFoot         102195
Head              54960
OtherBodyPart      1267
Name: shotType, dtype: int64
MissedShots    126980
BlockedShot     79992
SavedShot       75801
Goal            34498
ShotOnPost       6258
OwnGoal          1014
Name: shotResult, dtype: int64
```

```
0.885    5844
0.913    2425
0.910    2367
0.917    2351
0.919    2321
         ...
0.116       1
0.377       1
0.382       1
0.424       1
0.365       1
Name: positionX, Length: 813, dtype: int64
0.500    4738
0.534    1740
0.466    1730
0.493    1704
0.487    1698
         ...
0.127       1
0.915       1
0.989       1
0.058       1
0.133       1
Name: positionY, Length: 930, dtype: int64

count    324543.000000
mean          0.843968
std           0.090014
min           0.003000
25%           0.781000
50%           0.863000
75%           0.909000
max           0.999000
Name: positionX, dtype: float64
count    324543.000000
mean          0.504613
std           0.129372
min           0.000000
25%           0.414000
50%           0.501000
75%           0.597000
max           0.997000
Name: positionY, dtype: float64
```

| | DataFrame 1 | DataFrame 2 | Matching Features |
|---|---|---|---|
| **0** | leagues | teams | name |
| **1** | leagues | players | name |
| **2** | leagues | appearances | leagueID |
| **3** | leagues | games | leagueID |
| **4** | teamstats | teams | teamID |
| **5** | teamstats | shots | gameID |
| **6** | teamstats | appearances | shots, xGoals, goals, gameID |
| **7** | teamstats | games | season, date, gameID |
| **8** | teams | players | name |
| **9** | shots | appearances | gameID |
| **10** | shots | games | gameID |
| **11** | players | appearances | playerID |
| **12** | appearances | games | leagueID, gameID |

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals |
|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **12675** | 16131 | 5 | 2020 | 2021-05-23 19:00:00 | 168 | 166 | 1 | 2 |
| **12676** | 16132 | 5 | 2020 | 2021-05-23 19:00:00 | 177 | 176 | 1 | 2 |
| **12677** | 16133 | 5 | 2020 | 2021-05-23 19:00:00 | 163 | 235 | 2 | 0 |
| **12678** | 16134 | 5 | 2020 | 2021-05-23 19:00:00 | 175 | 181 | 0 | 1 |
| **12679** | 16135 | 5 | 2020 | 2021-05-23 19:00:00 | 225 | 179 | 1 | 1 |

12680 rows × 39 columns

Connecting playerID to teamID for further work on shots

| | gameID | teamID | playerID |
|---|---|---|---|
| **0** | 81 | 89 | 560 |
| **1** | 81 | 89 | 557 |
| **2** | 81 | 89 | 548 |
| **3** | 81 | 89 | 628 |
| **4** | 81 | 89 | 1006 |
| **...** | ... | ... | ... |
| **356508** | 16135 | 179 | 3509 |
| **356509** | 16135 | 179 | 4882 |
| **356510** | 16135 | 179 | 5786 |
| **356511** | 16135 | 179 | 8997 |
| **356512** | 16135 | 179 | 5762 |

356513 rows × 3 columns

| | gameID | teamID | playerID | playerName | teamName |
|---|---|---|---|---|---|
| **0** | 81 | 89 | 560 | Sergio Romero | Manchester United |
| **1** | 81 | 89 | 557 | Matteo Darmian | Manchester United |
| **2** | 81 | 89 | 548 | Daley Blind | Manchester United |
| **3** | 81 | 89 | 628 | Chris Smalling | Manchester United |
| **4** | 81 | 89 | 1006 | Luke Shaw | Manchester United |

```
No discrepancies found. Each player is assigned exactly one team per game.

None

No season-based team assignment violations found.
```

| | playerID | teamID | playerName | teamName |
|---|---|---|---|---|
| **0** | 560 | 89 | Sergio Romero | Manchester United |
| **1** | 557 | 89 | Matteo Darmian | Manchester United |
| **2** | 548 | 89 | Daley Blind | Manchester United |
| **3** | 628 | 89 | Chris Smalling | Manchester United |
| **4** | 1006 | 89 | Luke Shaw | Manchester United |
| **...** | ... | ... | ... | ... |
| **10101** | 7396 | 176 | Loic Bessile | Bordeaux |
| **10102** | 9566 | 175 | Yanis Lhéry | Saint-Etienne |
| **10103** | 9565 | 175 | Mathys Saban | Saint-Etienne |
| **10104** | 9568 | 181 | Charles Costes | Dijon |
| **10105** | 9567 | 181 | Erwan Belhadji | Dijon |

10106 rows × 4 columns

| | playerID | playerName | teams_played_for_names |
|---|---|---|---|
| **0** | 1 | Christian Mathenia | [Hamburger SV, Darmstadt, Nuernberg] |
| **1** | 2 | György Garics | [Darmstadt] |
| **2** | 3 | Luca Caldirola | [Werder Bremen, Darmstadt, Benevento] |
| **3** | 4 | Aytac Sulu | [Darmstadt] |
| **4** | 5 | Fabian Holland | [Darmstadt] |
| **5** | 6 | Marcel Heller | [Darmstadt, Augsburg] |
| **6** | 7 | Florian Jungwirth | [Darmstadt] |
| **7** | 8 | Jérôme Gondorf | [Werder Bremen, Darmstadt, Freiburg] |
| **8** | 9 | Tobias Kempe | [Darmstadt] |
| **9** | 10 | Jan Rosenthal | [Darmstadt] |

| | playerID | playerName | teams_played_for_names |
|---|---|---|---|
| **810** | 841 | Graziano Pellè | [Parma Calcio 1913, Southampton] |

creating player-appearance and shots df

| | gameID | teamID | total_assists | total_xAssists | total_key_passes | total_xGoalsChain | total_ |
|---|---|---|---|---|---|---|---|
| **0** | 81 | 82 | 0 | 0.586365 | 7 | 1.745371 | |
| **1** | 81 | 89 | 0 | 0.284979 | 5 | 1.396328 | |
| **2** | 82 | 71 | 1 | 0.560695 | 4 | 1.238205 | |
| **3** | 82 | 73 | 0 | 0.419975 | 9 | 2.159510 | |
| **4** | 83 | 72 | 2 | 0.549139 | 8 | 1.025550 | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **25355** | 16133 | 235 | 0 | 0.216965 | 6 | 0.884652 | |
| **25356** | 16134 | 175 | 0 | 1.265829 | 13 | 4.790546 | |
| **25357** | 16134 | 181 | 1 | 0.565077 | 6 | 1.256511 | |
| **25358** | 16135 | 179 | 1 | 0.470476 | 4 | 0.502347 | |
| **25359** | 16135 | 225 | 0 | 0.074636 | 4 | 0.528499 | |

25360 rows × 11 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25360 entries, 0 to 25359
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   gameID               25360 non-null  int64
 1   teamID               25360 non-null  int64
 2   total_assists        25360 non-null  int64
 3   total_xAssists       25360 non-null  float64
 4   total_key_passes     25360 non-null  int64
 5   total_xGoalsChain    25360 non-null  float64
 6   total_xGoalsBuildup  25360 non-null  float64
 7   total_yellow_cards   25360 non-null  int64
 8   total_red_cards      25360 non-null  int64
 9   total_blocked_shots  25349 non-null  float64
 10  total_saved_shots    25349 non-null  float64
dtypes: float64(5), int64(6)
memory usage: 2.1 MB
```

| | gameID | teamID | season | date | location | goals | xGoals | shots | shotsOnTarget | deep |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 89 | 2015 | 2015-08-08 15:45:00 | h | 1 | 0.627539 | 9 | 1 | 4 |
| **1** | 81 | 82 | 2015 | 2015-08-08 15:45:00 | a | 0 | 0.674600 | 9 | 4 | 10 |
| **2** | 82 | 73 | 2015 | 2015-08-08 18:00:00 | h | 0 | 0.876106 | 11 | 2 | 11 |
| **3** | 82 | 71 | 2015 | 2015-08-08 18:00:00 | a | 1 | 0.782253 | 7 | 3 | 2 |
| **4** | 83 | 72 | 2015 | 2015-08-08 18:00:00 | h | 2 | 0.604226 | 10 | 5 | 5 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | .. |
| **25355** | 16133 | 235 | 2020 | 2021-05-23 19:00:00 | a | 0 | 0.357583 | 9 | 2 | 0 |
| **25356** | 16134 | 175 | 2020 | 2021-05-23 19:00:00 | h | 0 | 1.460500 | 19 | 5 | 6 |
| **25357** | 16134 | 181 | 2020 | 2021-05-23 19:00:00 | a | 1 | 1.380290 | 10 | 2 | 3 |
| **25358** | 16135 | 225 | 2020 | 2021-05-23 19:00:00 | h | 1 | 0.323960 | 6 | 2 | 1 |
| **25359** | 16135 | 179 | 2020 | 2021-05-23 19:00:00 | a | 1 | 0.521913 | 7 | 1 | 0 |

25360 rows × 25 columns

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 25360 entries, 0 to 25359
Data columns (total 25 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   gameID               25360 non-null  int64
 1   teamID               25360 non-null  int64
 2   season               25360 non-null  int64
 3   date                 25360 non-null  object
 4   location             25360 non-null  object
 5   goals                25360 non-null  int64
 6   xGoals               25360 non-null  float64
 7   shots                25360 non-null  int64
 8   shotsOnTarget        25360 non-null  int64
 9   deep                 25360 non-null  int64
 10  ppda                 25360 non-null  float64
 11  fouls                25360 non-null  int64
 12  corners              25360 non-null  int64
 13  yellowCards          25359 non-null  float64
 14  redCards             25360 non-null  int64
 15  result               25360 non-null  object
 16  total_assists        25360 non-null  int64
 17  total_xAssists       25360 non-null  float64
 18  total_key_passes     25360 non-null  int64
 19  total_xGoalsChain    25360 non-null  float64
 20  total_xGoalsBuildup  25360 non-null  float64
 21  total_yellow_cards   25360 non-null  int64
 22  total_red_cards      25360 non-null  int64
 23  total_blocked_shots  25349 non-null  float64
 24  total_saved_shots    25349 non-null  float64
dtypes: float64(8), int64(14), object(3)
memory usage: 5.0+ MB
None

0        W
1        L
2        L
3        W
4        D
        ..
25355    L
25356    L
25357    W
25358    D
25359    D
Name: result, Length: 25360, dtype: object
```

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals |
|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **12675** | 16131 | 5 | 2020 | 2021-05-23 19:00:00 | 168 | 166 | 1 | 2 |
| **12676** | 16132 | 5 | 2020 | 2021-05-23 19:00:00 | 177 | 176 | 1 | 2 |
| **12677** | 16133 | 5 | 2020 | 2021-05-23 19:00:00 | 163 | 235 | 2 | 0 |
| **12678** | 16134 | 5 | 2020 | 2021-05-23 19:00:00 | 175 | 181 | 0 | 1 |
| **12679** | 16135 | 5 | 2020 | 2021-05-23 19:00:00 | 225 | 179 | 1 | 1 |

12680 rows × 55 columns

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 55 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gameID                    12680 non-null  int64
 1   leagueID                  12680 non-null  int64
 2   season                    12680 non-null  int64
 3   date                      12680 non-null  object
 4   homeTeamID                12680 non-null  int64
 5   awayTeamID                12680 non-null  int64
 6   homeGoals                 12680 non-null  int64
 7   awayGoals                 12680 non-null  int64
 8   homeGoalsHalfTime         12680 non-null  int64
 9   awayGoalsHalfTime         12680 non-null  int64
 10  home_season               12680 non-null  int64
 11  home_date                 12680 non-null  object
 12  home_location             12680 non-null  object
 13  home_xGoals               12680 non-null  float64
 14  home_shots                12680 non-null  int64
 15  home_shotsOnTarget        12680 non-null  int64
 16  home_deep                 12680 non-null  int64
 17  home_ppda                 12680 non-null  float64
 18  home_fouls                12680 non-null  int64
 19  home_corners              12680 non-null  int64
 20  home_yellowCards          12679 non-null  float64
 21  home_redCards             12680 non-null  int64
 22  home_result               12680 non-null  object
 23  home_total_assists        12680 non-null  int64
 24  home_total_xAssists       12680 non-null  float64
 25  home_total_key_passes     12680 non-null  int64
 26  home_total_xGoalsChain    12680 non-null  float64
 27  home_total_xGoalsBuildup  12680 non-null  float64
 28  home_total_yellow_cards   12680 non-null  int64
 29  home_total_red_cards      12680 non-null  int64
 30  home_total_blocked_shots  12677 non-null  float64
 31  home_total_saved_shots    12677 non-null  float64
 32  away_season               12680 non-null  int64
 33  away_date                 12680 non-null  object
 34  away_location             12680 non-null  object
 35  away_xGoals               12680 non-null  float64
 36  away_shots                12680 non-null  int64
 37  away_shotsOnTarget        12680 non-null  int64
 38  away_deep                 12680 non-null  int64
 39  away_ppda                 12680 non-null  float64
 40  away_fouls                12680 non-null  int64
 41  away_corners              12680 non-null  int64
 42  away_yellowCards          12680 non-null  float64
 43  away_redCards             12680 non-null  int64
 44  away_result               12680 non-null  object
 45  away_total_assists        12680 non-null  int64
 46  away_total_xAssists       12680 non-null  float64
 47  away_total_key_passes     12680 non-null  int64
 48  away_total_xGoalsChain    12680 non-null  float64
 49  away_total_xGoalsBuildup  12680 non-null  float64
 50  away_total_yellow_cards   12680 non-null  int64
 51  away_total_red_cards      12680 non-null  int64
 52  away_total_blocked_shots  12672 non-null  float64
 53  away_total_saved_shots    12672 non-null  float64
```

```
 54  gameresult                12680 non-null  object
dtypes: float64(16), int64(31), object(8)
memory usage: 5.4+ MB


<class 'pandas.core.frame.DataFrame'>
Int64Index: 12680 entries, 0 to 12679
Data columns (total 47 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gameID                    12680 non-null  int64
 1   leagueID                  12680 non-null  int64
 2   season                    12680 non-null  int64
 3   date                      12680 non-null  object
 4   homeTeamID                12680 non-null  int64
 5   awayTeamID                12680 non-null  int64
 6   homeGoals                 12680 non-null  int64
 7   awayGoals                 12680 non-null  int64
 8   homeGoalsHalfTime         12680 non-null  int64
 9   awayGoalsHalfTime         12680 non-null  int64
 10  home_xGoals               12680 non-null  float64
 11  home_shots                12680 non-null  int64
 12  home_shotsOnTarget        12680 non-null  int64
 13  home_deep                 12680 non-null  int64
 14  home_ppda                 12680 non-null  float64
 15  home_fouls                12680 non-null  int64
 16  home_corners              12680 non-null  int64
 17  home_yellowCards          12679 non-null  float64
 18  home_redCards             12680 non-null  int64
 19  home_total_assists        12680 non-null  int64
 20  home_total_xAssists       12680 non-null  float64
 21  home_total_key_passes     12680 non-null  int64
 22  home_total_xGoalsChain    12680 non-null  float64
 23  home_total_xGoalsBuildup  12680 non-null  float64
 24  home_total_yellow_cards   12680 non-null  int64
 25  home_total_red_cards      12680 non-null  int64
 26  home_total_blocked_shots  12677 non-null  float64
 27  home_total_saved_shots    12677 non-null  float64
 28  away_xGoals               12680 non-null  float64
 29  away_shots                12680 non-null  int64
 30  away_shotsOnTarget        12680 non-null  int64
 31  away_deep                 12680 non-null  int64
 32  away_ppda                 12680 non-null  float64
 33  away_fouls                12680 non-null  int64
 34  away_corners              12680 non-null  int64
 35  away_yellowCards          12680 non-null  float64
 36  away_redCards             12680 non-null  int64
 37  away_total_assists        12680 non-null  int64
 38  away_total_xAssists       12680 non-null  float64
 39  away_total_key_passes     12680 non-null  int64
 40  away_total_xGoalsChain    12680 non-null  float64
 41  away_total_xGoalsBuildup  12680 non-null  float64
 42  away_total_yellow_cards   12680 non-null  int64
 43  away_total_red_cards      12680 non-null  int64
 44  away_total_blocked_shots  12672 non-null  float64
 45  away_total_saved_shots    12672 non-null  float64
 46  gameresult                12680 non-null  object
dtypes: float64(16), int64(29), object(2)
memory usage: 4.6+ MB
```

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals |
|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **12675** | 16131 | 5 | 2020 | 2021-05-23 19:00:00 | 168 | 166 | 1 | 2 |
| **12676** | 16132 | 5 | 2020 | 2021-05-23 19:00:00 | 177 | 176 | 1 | 2 |
| **12677** | 16133 | 5 | 2020 | 2021-05-23 19:00:00 | 163 | 235 | 2 | 0 |
| **12678** | 16134 | 5 | 2020 | 2021-05-23 19:00:00 | 175 | 181 | 0 | 1 |
| **12679** | 16135 | 5 | 2020 | 2021-05-23 19:00:00 | 225 | 179 | 1 | 1 |

12680 rows × 47 columns

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals |
|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **12675** | 16131 | 5 | 2020 | 2021-05-23 19:00:00 | 168 | 166 | 1 | 2 |
| **12676** | 16132 | 5 | 2020 | 2021-05-23 19:00:00 | 177 | 176 | 1 | 2 |
| **12677** | 16133 | 5 | 2020 | 2021-05-23 19:00:00 | 163 | 235 | 2 | 0 |
| **12678** | 16134 | 5 | 2020 | 2021-05-23 19:00:00 | 175 | 181 | 0 | 1 |
| **12679** | 16135 | 5 | 2020 | 2021-05-23 19:00:00 | 225 | 179 | 1 | 1 |

12680 rows × 47 columns

create a dictionary of all the dataframes in use

```
=== df_appearances ===
Shape: (356513, 18)
```

| | gameID | playerID | goals | ownGoals | shots | xGoals | xGoalsChain | xGoalsBuildup | assists |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 560 | 0 | 0 | 0 | 0.0 | 0.000000 | 0.000000 | 0 |
| **1** | 81 | 557 | 0 | 0 | 0 | 0.0 | 0.106513 | 0.106513 | 0 |
| **2** | 81 | 548 | 0 | 0 | 0 | 0.0 | 0.127738 | 0.127738 | 0 |
| **3** | 81 | 628 | 0 | 0 | 0 | 0.0 | 0.106513 | 0.106513 | 0 |
| **4** | 81 | 1006 | 0 | 0 | 0 | 0.0 | 0.021225 | 0.021225 | 0 |

```
=== df_games ===
Shape: (12680, 10)
```

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals | home |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 | |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 | |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 | |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 | |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 | |

```
=== df_games_odds ===
Shape: (12680, 25)
```

| | gameID | homeProbability | drawProbability | awayProbability | B365H | B365D | B365A | BWH | BW |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 0.2843 | 0.3999 | 0.3158 | 1.65 | 4.0 | 6.00 | 1.65 | 4 |
| **1** | 82 | 0.3574 | 0.3500 | 0.2926 | 2.00 | 3.6 | 4.00 | 2.00 | 3 |
| **2** | 83 | 0.2988 | 0.4337 | 0.2675 | 1.70 | 3.9 | 5.50 | 1.70 | 3 |
| **3** | 84 | 0.6422 | 0.2057 | 0.1521 | 1.95 | 3.5 | 4.33 | 2.00 | 3 |
| **4** | 85 | 0.1461 | 0.2159 | 0.6380 | 2.55 | 3.3 | 3.00 | 2.60 | 3 |

5 rows × 25 columns

```
=== df_shots ===
Shape: (324543, 11)
```

| | gameID | shooterID | assisterID | minute | situation | lastAction | shotType | shotResult | |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 554 | NaN | 27 | DirectFreekick | Standard | LeftFoot | BlockedShot | 0. |
| **1** | 81 | 555 | 631.0 | 27 | SetPiece | Pass | RightFoot | BlockedShot | 0. |
| **2** | 81 | 554 | 629.0 | 35 | OpenPlay | Pass | LeftFoot | BlockedShot | 0. |
| **3** | 81 | 554 | NaN | 35 | OpenPlay | Tackle | LeftFoot | MissedShots | 0. |
| **4** | 81 | 555 | 654.0 | 40 | OpenPlay | BallRecovery | RightFoot | BlockedShot | 0. |

```
=== df_teamstats ===
Shape: (25360, 16)
```

| | gameID | teamID | season | date | location | goals | xGoals | shots | shotsOnTarget | deep | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 89 | 2015 | 2015-08-08 15:45:00 | h | 1 | 0.627539 | 9 | 1 | 4 | 13 |
| **1** | 81 | 82 | 2015 | 2015-08-08 15:45:00 | a | 0 | 0.674600 | 9 | 4 | 10 | 8 |
| **2** | 82 | 73 | 2015 | 2015-08-08 18:00:00 | h | 0 | 0.876106 | 11 | 2 | 11 | 6 |
| **3** | 82 | 71 | 2015 | 2015-08-08 18:00:00 | a | 1 | 0.782253 | 7 | 3 | 2 | 11 |
| **4** | 83 | 72 | 2015 | 2015-08-08 18:00:00 | h | 2 | 0.604226 | 10 | 5 | 5 | 6 |

```
=== df_combined ===
Shape: (12680, 39)
```

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals | home |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 | |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 | |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 | |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 | |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 | |

5 rows × 39 columns

```
=== player_game_team_mapping ===
Shape: (356513, 5)
```

| | gameID | teamID | playerID | playerName | teamName |
|---|---|---|---|---|---|
| **0** | 81 | 89 | 560 | Sergio Romero | Manchester United |
| **1** | 81 | 89 | 557 | Matteo Darmian | Manchester United |
| **2** | 81 | 89 | 548 | Daley Blind | Manchester United |
| **3** | 81 | 89 | 628 | Chris Smalling | Manchester United |
| **4** | 81 | 89 | 1006 | Luke Shaw | Manchester United |

```
=== player_shots ===
Shape: (324543, 11)
```

| | gameID | playerID | assisterID | minute | situation | lastAction | shotType | shotResult | x |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 554 | NaN | 27 | DirectFreekick | Standard | LeftFoot | BlockedShot | 0.10 |
| **1** | 81 | 555 | 631.0 | 27 | SetPiece | Pass | RightFoot | BlockedShot | 0.06 |
| **2** | 81 | 554 | 629.0 | 35 | OpenPlay | Pass | LeftFoot | BlockedShot | 0.05 |
| **3** | 81 | 554 | NaN | 35 | OpenPlay | Tackle | LeftFoot | MissedShots | 0.09 |
| **4** | 81 | 555 | 654.0 | 40 | OpenPlay | BallRecovery | RightFoot | BlockedShot | 0.03 |

```
=== player_performance ===
Shape: (25349, 2)
```

| | | total_blocked_shots | total_saved_shots |
|---|---|---|---|
| **gameID** | **teamID** | | |
| **81** | **82** | 3 | 4 |
| | **89** | 4 | 1 |
| **82** | **71** | 2 | 2 |
| | **73** | 2 | 2 |
| **83** | **72** | 2 | 3 |

```
=== team_performance ===
Shape: (25360, 7)
```

| | | total_assists | total_xAssists | total_key_passes | total_xGoalsChain | total_xGoalsB |
|---|---|---|---|---|---|---|
| **gameID** | **teamID** | | | | | |
| **81** | **82** | 0 | 0.586365 | 7 | 1.745371 | 0.8 |
| | **89** | 0 | 0.284979 | 5 | 1.396328 | 0.9 |
| **82** | **71** | 1 | 0.560695 | 4 | 1.238205 | 0.7 |
| | **73** | 0 | 0.419975 | 9 | 2.159510 | 1.1 |
| **83** | **72** | 2 | 0.549139 | 8 | 1.025550 | 0.4 |

```
=== teamstats ===
Shape: (12680, 47)
```

| | gameID | leagueID | season | date | homeTeamID | awayTeamID | homeGoals | awayGoals | home |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 81 | 1 | 2015 | 2015-08-08 15:45:00 | 89 | 82 | 1 | 0 | |
| **1** | 82 | 1 | 2015 | 2015-08-08 18:00:00 | 73 | 71 | 0 | 1 | |
| **2** | 83 | 1 | 2015 | 2015-08-08 18:00:00 | 72 | 90 | 2 | 2 | |
| **3** | 84 | 1 | 2015 | 2015-08-08 18:00:00 | 75 | 77 | 4 | 2 | |
| **4** | 85 | 1 | 2015 | 2015-08-08 18:00:00 | 79 | 78 | 1 | 3 | |

5 rows × 47 columns

# DATA PROTOCOL

```
Exported data protocol files for df_appearances
Exported data protocol files for df_games
Exported data protocol files for df_games_odds
Exported data protocol files for df_shots
Exported data protocol files for df_teamstats
Exported data protocol files for df_combined
Exported data protocol files for player_game_team_mapping
Exported data protocol files for player_shots
Exported data protocol files for player_performance
Exported data protocol files for team_performance
Exported data protocol files for teamstats
```

```
Saved df_appearances to ../pickles/df_appearances.pkl
Saved df_games to ../pickles/df_games.pkl
Saved df_games_odds to ../pickles/df_games_odds.pkl
Saved df_shots to ../pickles/df_shots.pkl
Saved df_teamstats to ../pickles/df_teamstats.pkl
Saved df_combined to ../pickles/df_combined.pkl
Saved player_game_team_mapping to ../pickles/player_game_team_mapping.pkl
Saved player_shots to ../pickles/player_shots.pkl
Saved player_performance to ../pickles/player_performance.pkl
Saved team_performance to ../pickles/team_performance.pkl
Saved teamstats to ../pickles/teamstats.pkl
```